VA | U.S. Department of Veterans Affairs

Exploratory analysis:
Understanding trends and patterns

Platform Analytics & Insights Training

Chloe Brown, Data Analyst
chloe.brown@adhocteam.us

October 2021

Goal

Provide VFS teams with actionable steps to:

- Explore and visualize data with purpose
- Ask questions and interpret results
- Make decisions about what to do next

This training assumes you already have:

- Well-defined KPIs for your product
- Understanding of how data is collected
- Understanding of what variables mean
- Verified data quality/cleanliness

Scope

Today's focus

This training will mostly focus on **descriptive statistics**. This will give us a foundation to build up to more complex analysis next.

Stay tuned!

Future Platform Analytics trainings will cover:

- Assessing data quality
- A/B testing
- Experiment design

Agenda

Univariate analysis

One thing at a time

Frequency distributions

Measures of center

Measures of spread

Bivariate analysis

Two things at a time

Visualizing data

Describing patterns

Understanding correlation

Univariate analysis

Definition

Univariate analysis explores each of the variables in your data separately.

This gives a sense of how the variable behaves before you begin looking at relationships or how things change over time.

Exploratory tools:

- Frequency distributions
- Measures of center
- Measures of spread

Frequency distributions

What it means

How often each category or value appears in the data

Why it matters

“30,000 view” of data points for a variable

What to look at

tables, bar charts, histograms

Frequency distributions

Frequency table

Strengths:

- Easy to see specific totals
- More readable way to view many outputs at once

Top referral sources

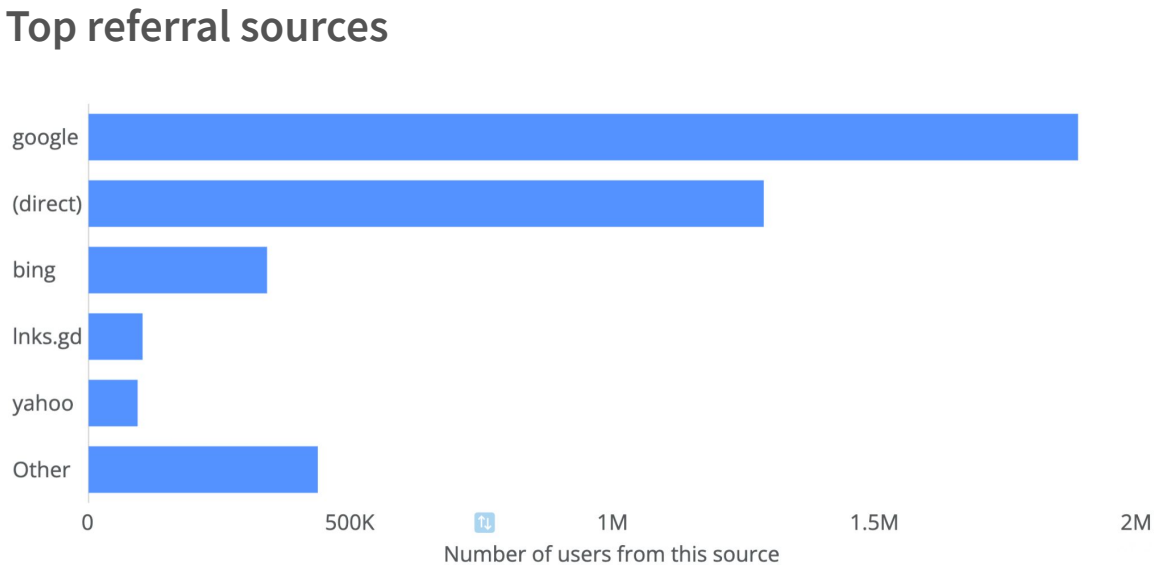
Source	User count
google	1,887,709
(direct)	1,288,172
bing	340,545
lnks.gd	103,031
yahoo	93,590
search.usa.gov	45,455
duckduckgo	23,332
links.govdelivery.com	16,610
VANotify	16,144
military.com	14,235
m.facebook.com	13,323
osd.mil	11,344
id.me	10,888
Newsletter	8,304
usa.gov	6,895
public.govdelivery.com	6,272

Frequency distributions

Bar chart

Strengths:

- Easy to understand relative totals at a glance
- Good for categorical or discrete numerical data

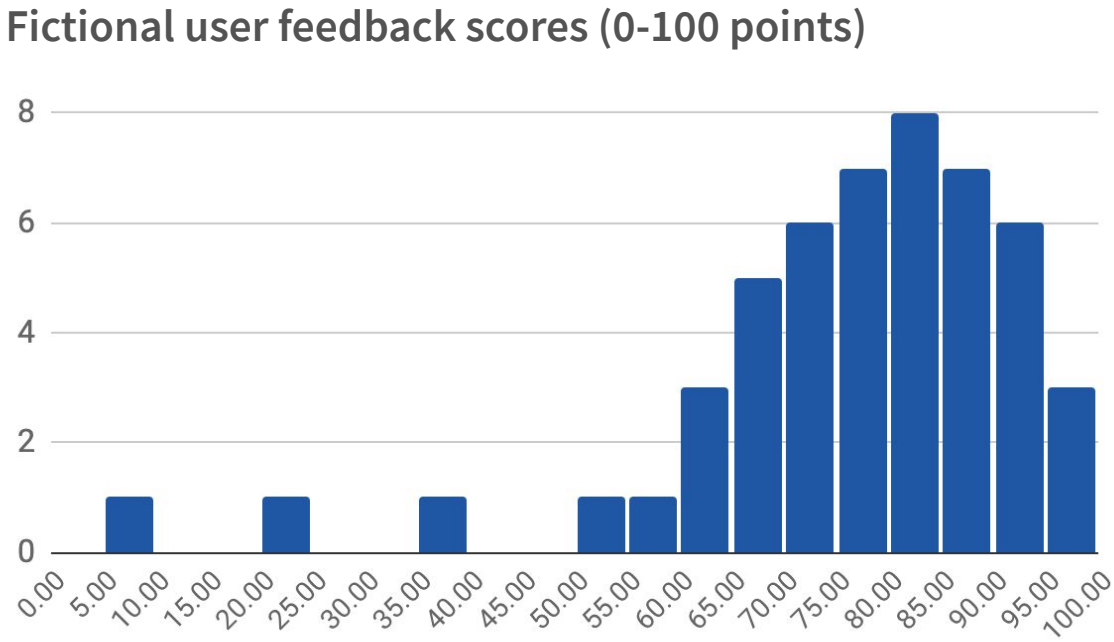


Frequency distributions

Histogram

Strengths:

- Good for continuous data, since they are grouped by bins
- Helps visualize
  - Skew
  - Tails



Measures of center

What it means

Different ways to define the “middle” for your data points

Why it matters

Estimates a “typical” value for the variable

What to look at

mean, median, mode

Measures of center

Mean

Sum of all data points divided by the number of data points.

Mean is the most commonly used measure of center, often called the average.

- **Pro:** easy to calculate
- **Con:** affected by outliers

Avg. time to complete a process

Avg. attendees per event in a series

Measures of center

Median

Middle value when data points are sorted (or mean of middle two, if an even number).

Less commonly used, except for data with extreme outliers.

- **Pro:** unaffected by outliers
- **Con:** may be more difficult to calculate (or explain)

Median income in a city or state

Median home price in a zip code

Measures of center

Mode

Most frequently occurring value for the variable.

Some data may be multi-modal (more than one mode value) or have no mode at all, depending on distribution.

- **Pro:** easy to calculate
- **Con:** limited use value

Most frequent customer rating

Most common cities where users live

Measures of spread

What it means

How close together or far apart the data points are from each other

Why it matters

Gives context for interpreting measures of center

What to look at

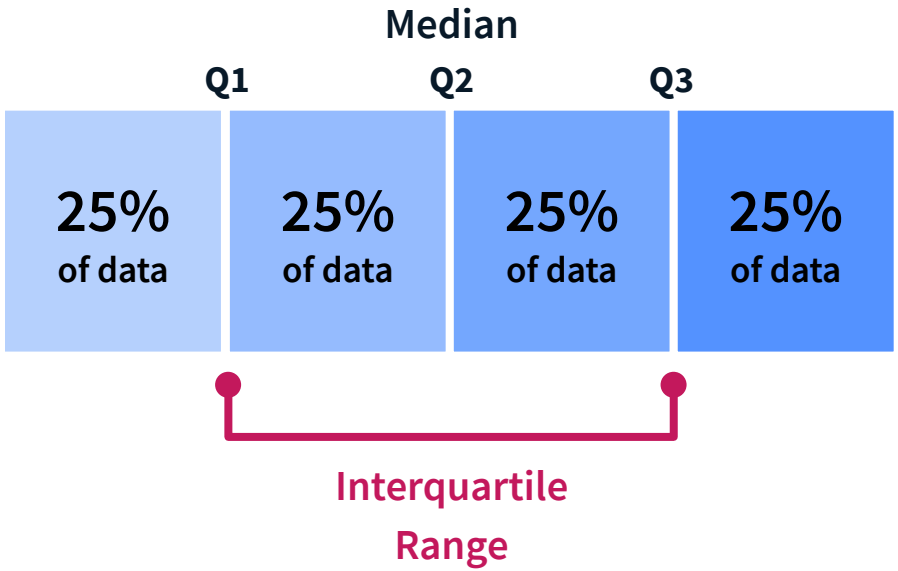
interquartile range, standard deviation

Measures of spread

Interquartile range (IQR)

Difference between the 1st and 3rd quartile variables, when all data points are ordered. This range describes the middle 50% of data points.

- Use as the measure of spread when using the median
- Use to calculate outliers in univariate analysis:
  - < Q1 – (1.5 * IQR)
  - > Q3 + (1.5 * IQR)



Measures of spread

Standard deviation

Technical definition: square root of the variance, which is the collective average of all squared differences between each data point and the mean.

When a distribution has been normalized, we can assume that a certain percentage of the observations fall between 1, 3, or 3 standard deviations from the mean.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

s = sample standard deviation

N = the number of observations

x_i = the observed values of a sample item

$\overline{x}$ = the mean value of the observations

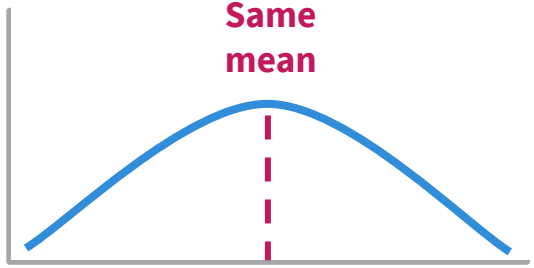
Measures of spread

Standard deviation

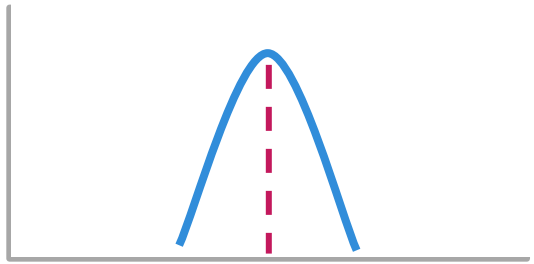
Plain language: The most common way of describing dispersion of a set of data from its mean. The *smaller* the standard deviation is, the more tightly clustered around the center the data will be.

It'll also become important when we talk about testing in future trainings.

High standard deviation



Low standard deviation



Bivariate analysis

Definition

Bivariate analysis plots two quantitative variables against each other.

Doing this visually will give you a sense of shape, direction, and strength of any relationship between your variables.

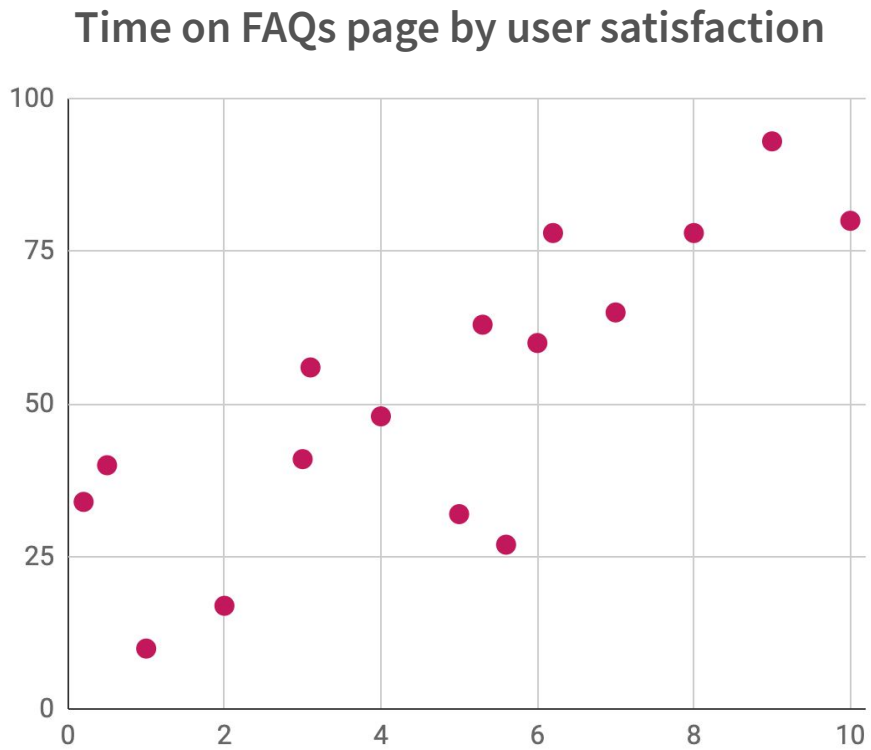
Exploratory tools:

- Visualizing your data
- Common patterns
- Correlation

Visualizing data

Composing scatter plots

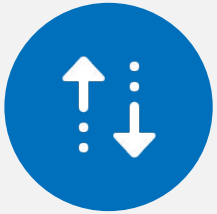
Set up using two numeric variables with a dot representing each data point.



Describing patterns

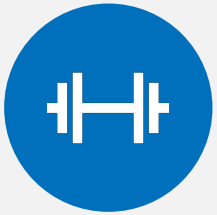
Shape

Linear or nonlinear? Would a straight line best describe the pattern you see?



Direction

Positive or negative? Does the relationship seem to be trending up or down?



Strength

Strong or weak? Are most dots clustered near the line or farther away?

Outliers

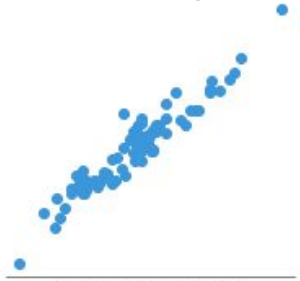
Which data points break the pattern? More of an art than a science.

Describing patterns

Examples:

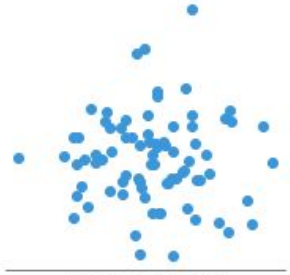
Shape
Direction
Strength
Outliers

Linear, positive, strong

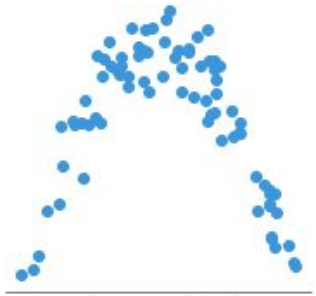


Linear, negative, moderate

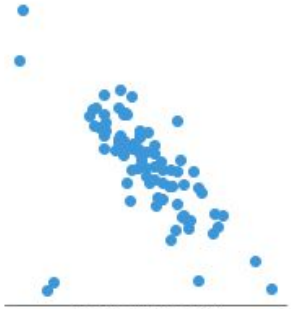
No relationship



Non-linear, strong



Outliers



Graphics courtesy of Chartio.

Describing patterns

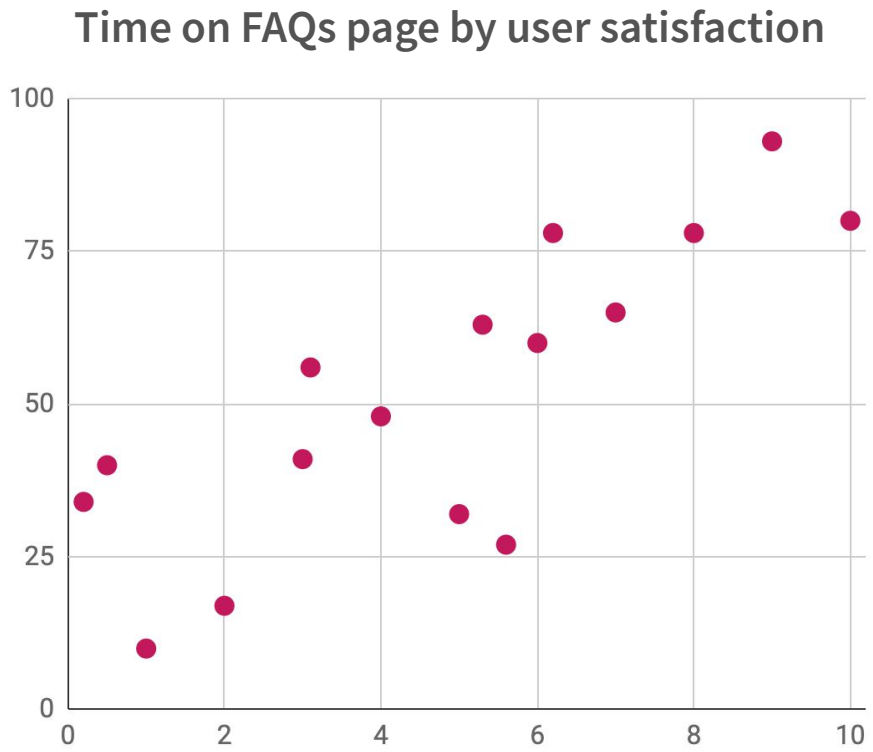
Returning to our example

Shape: Linear

Direction: Positive

Strength: Moderate

Outliers: n/a



Correlation

Definition

The strength of a **linear relationship** between two quantitative variables.

How it's measured

- Pearson's correlation coefficient, commonly shown as r = X.
- Can be any value between –1 (perfect negative relationship) and 1 (perfect positive relationship). A coefficient of 0 represents no relationship between variables.

Correlation

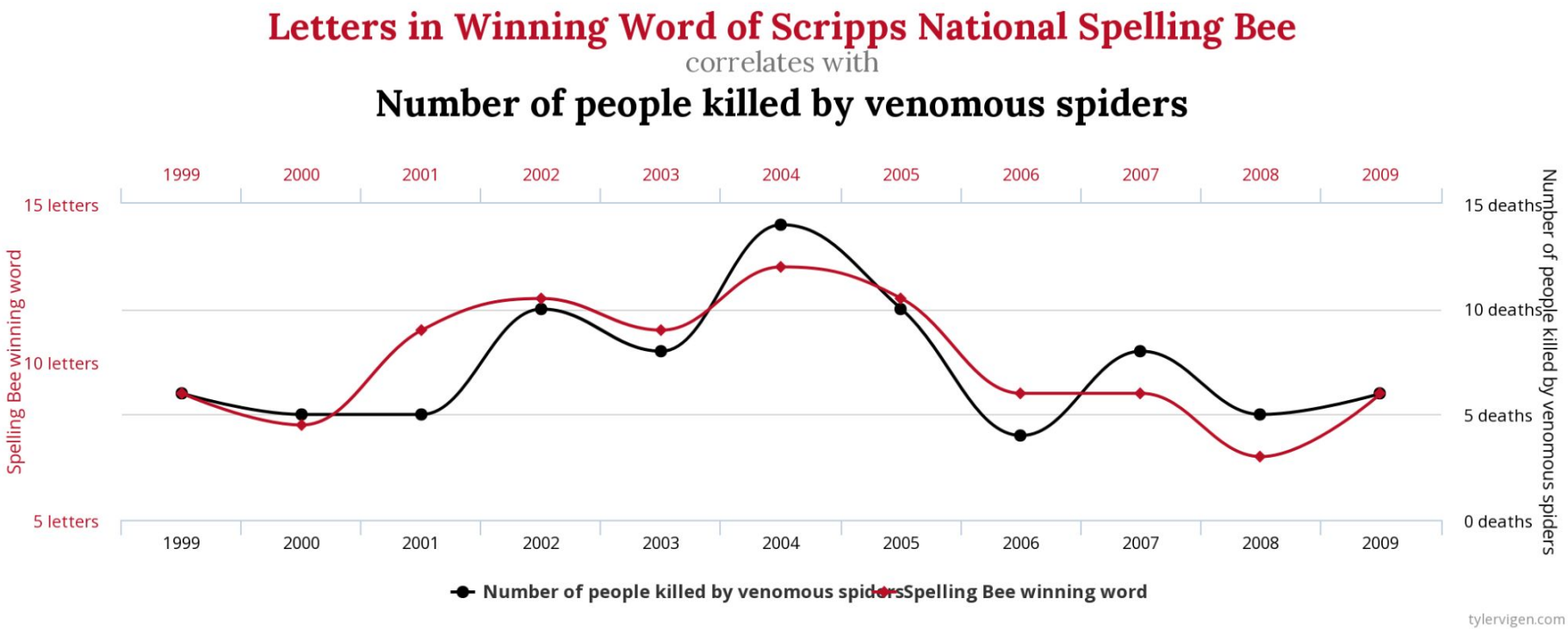
Correlation famously does not equal causation. But why not?

- The pattern you're observing could be completely random.
- The causal relationship could be reversed.
- There could be a third, unseen variable affecting both.

Sounds easy! But it can be more challenging in practice, especially when stakeholders are involved.

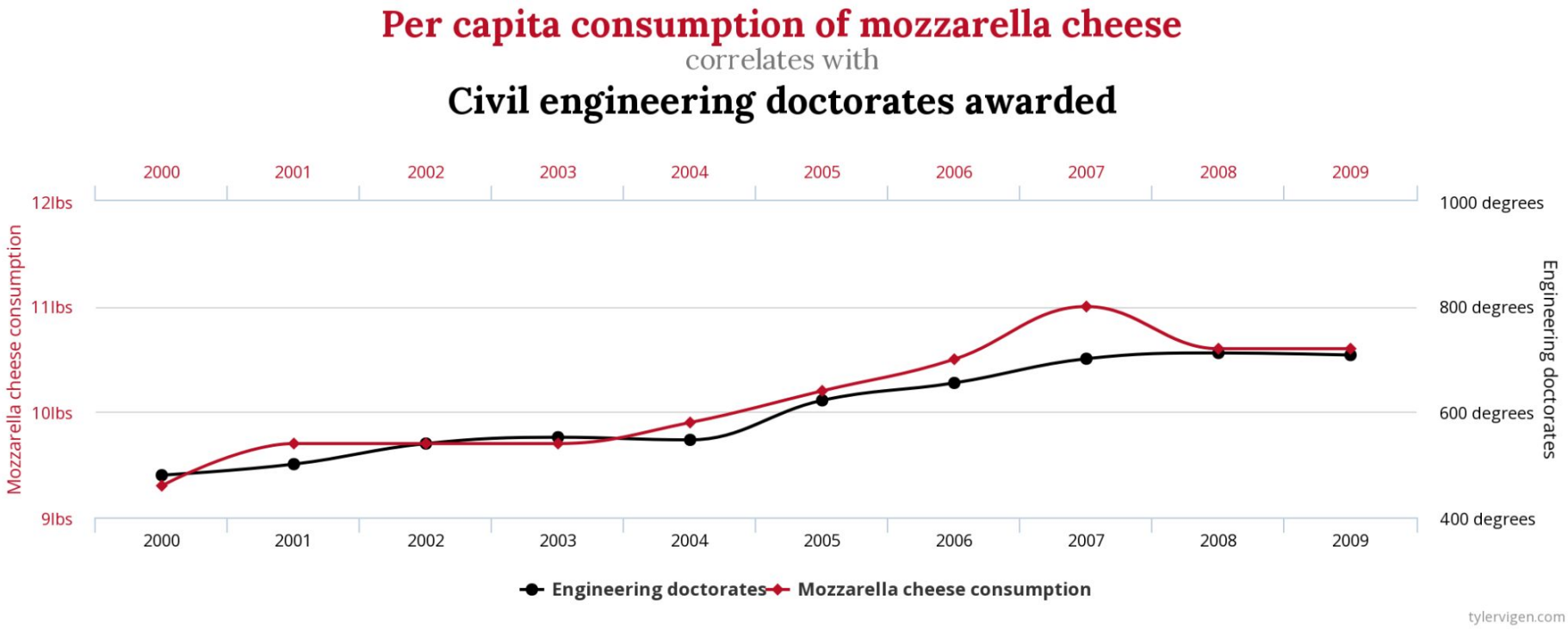
Spurious correlations

r = 0.8057



Spurious correlations

r = 0.9586



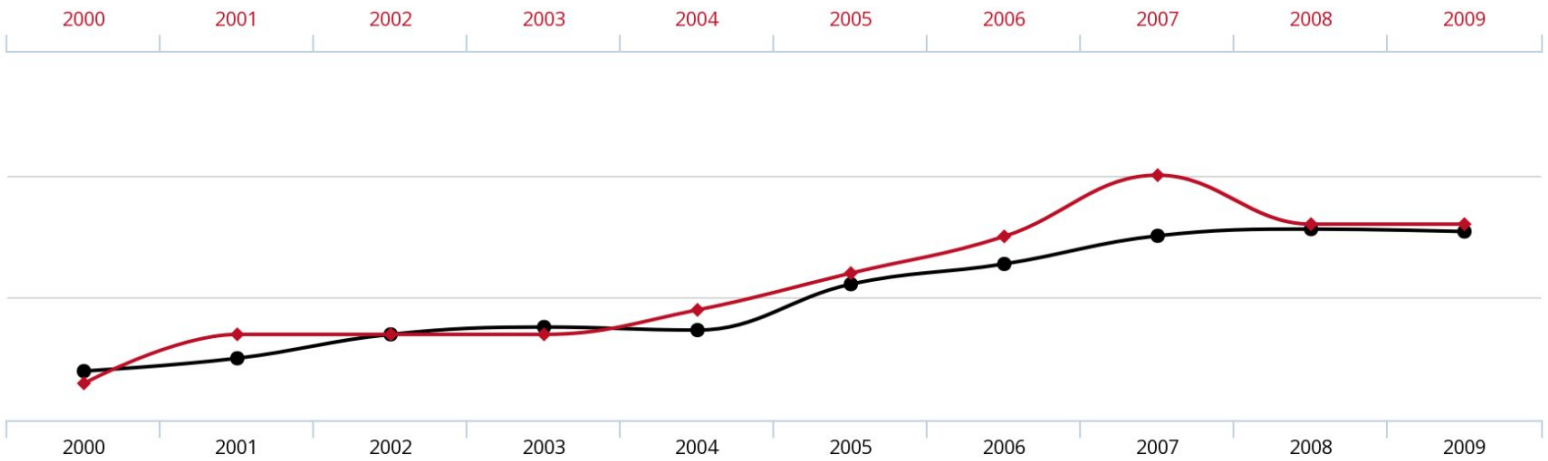
Spurious correlations

r = 0.9586

The thing I have control over

correlates with

The thing I want to impact



Correlation

We are all susceptible to confirmation bias.

What are the appropriate ways to talk about correlations when we find them?

✔ *These variables seem to have a strong relationship / association.*

✔ *We're interested in pursuing deeper analysis into X and Y, given their strong correlation.*

✘ *Changes in X are causing outcome Y, given their strong correlation.*

What's next?

To understand whether a relationship between two variables is causal, you'll need to test it.

These tools will be the topic of our next Insights trainings.

Have questions or interested in providing input on what future trainings should cover? Reach out to the Analytics team on #vfs-platform-support on Slack.

Q&A

Icons courtesy of icons8.